DSP543 Final Report

Jiayuan Zhang

May 10, 2019

Empirical investigation on peple's purchasing behavior on black friday

Data Description

Figure 1 Count of age group

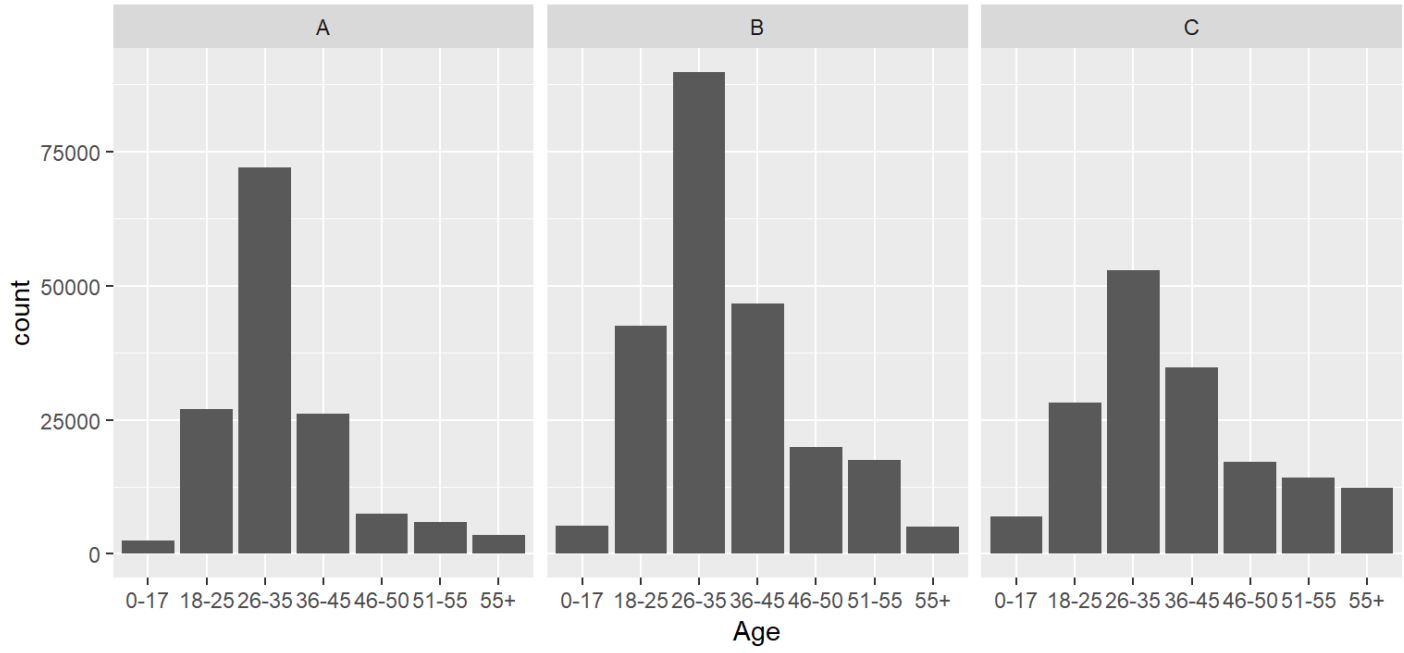


Figure 2 Count of gender group

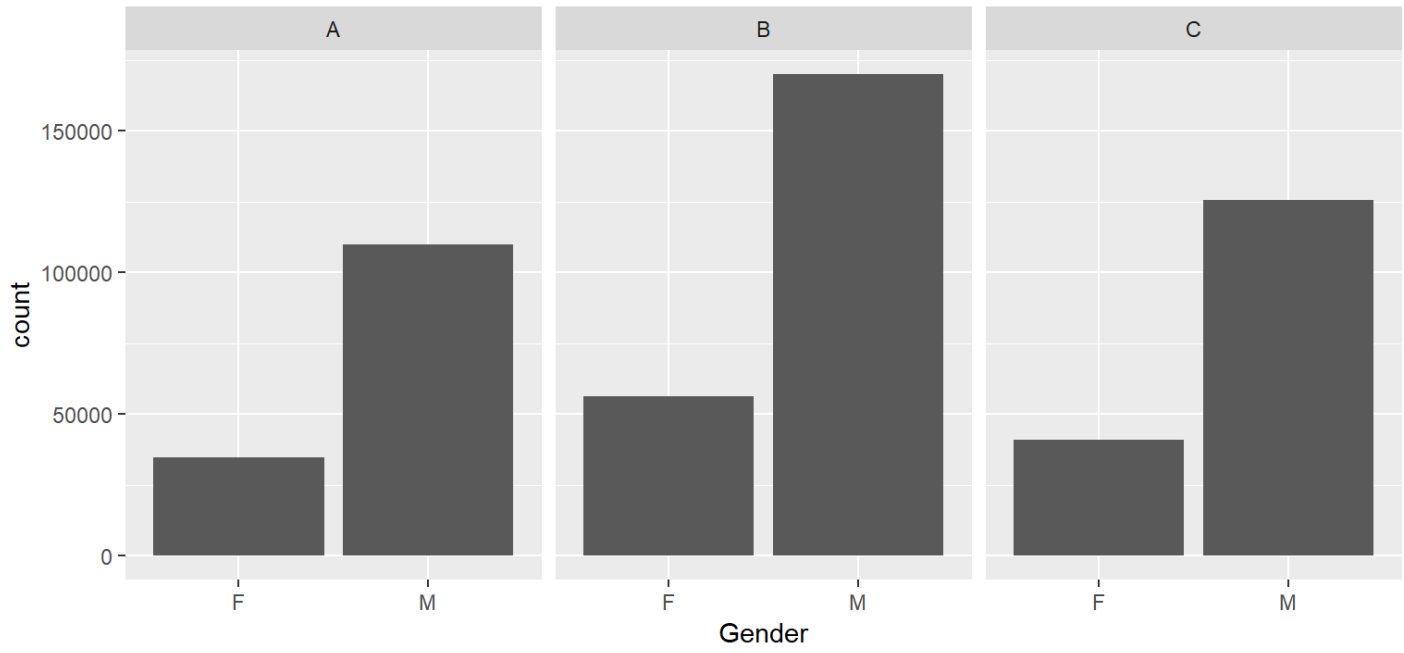


Figure 3 Count of occupation group

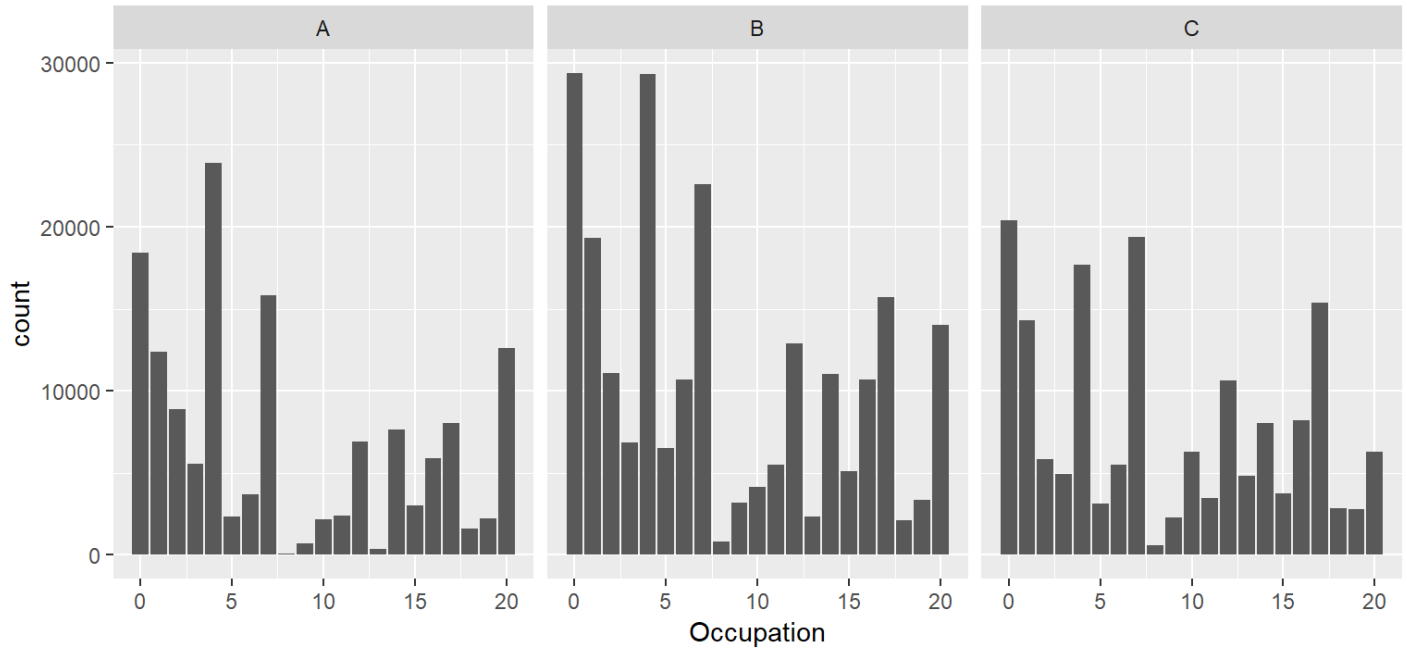


Figure 4 Count of marital status

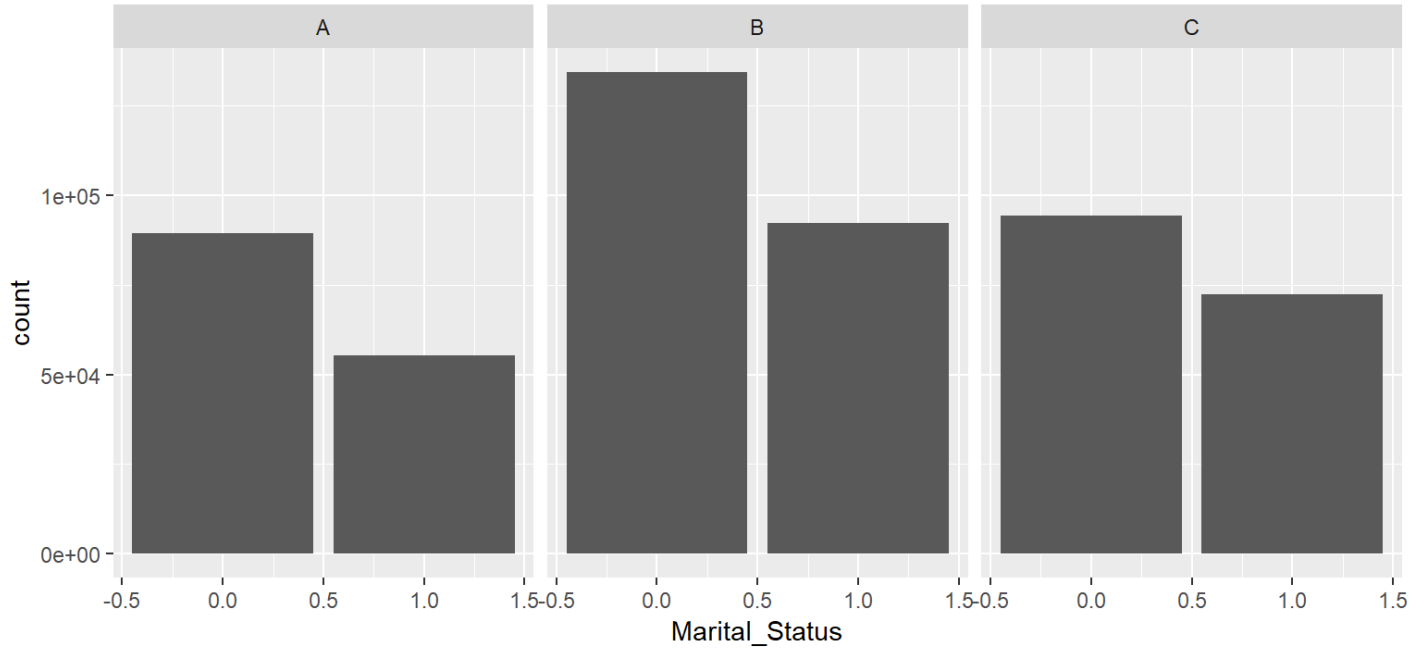


Figure 5 - Boxplot of purchase amount

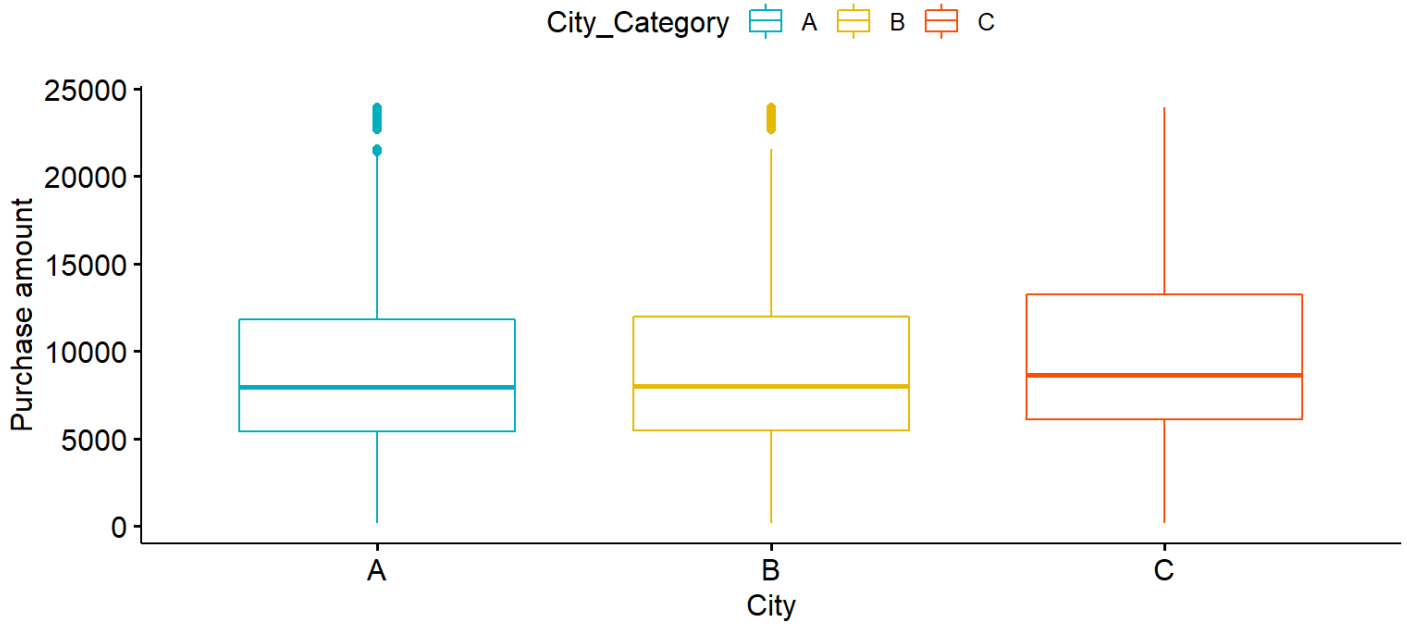


Figure 6 - Boxplot of items in category 1

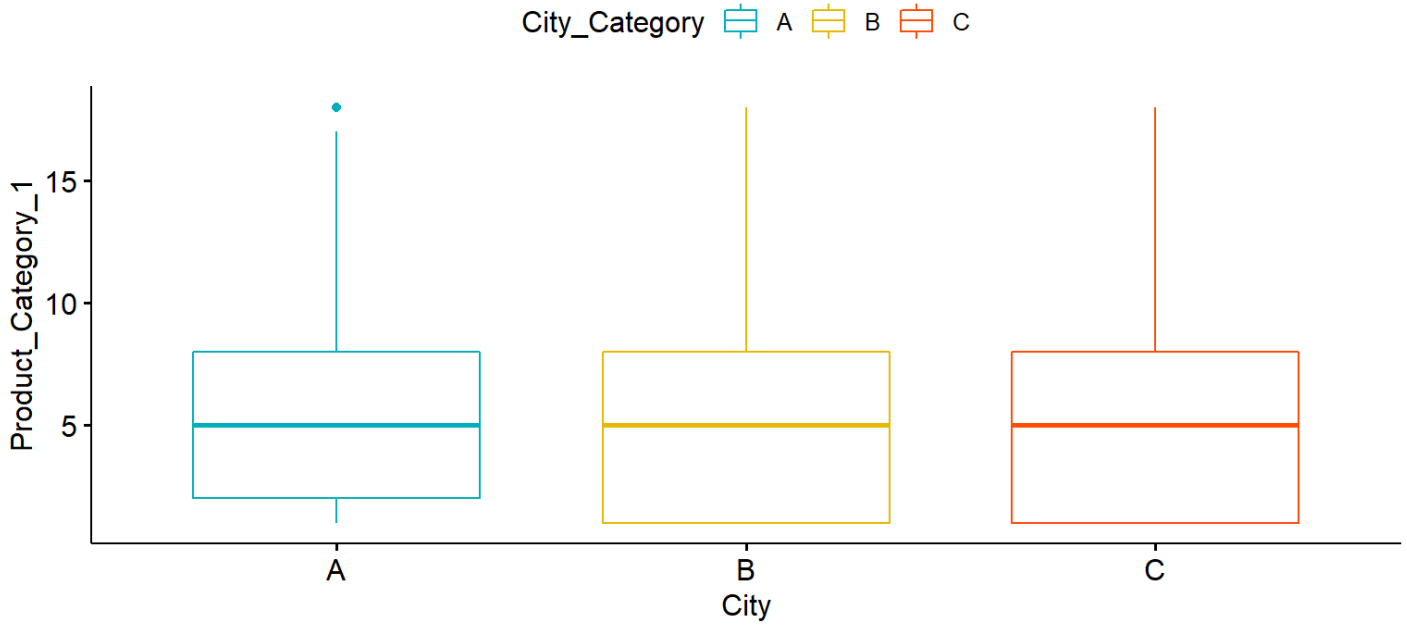


Figure 7 - Boxplot of items in category 2

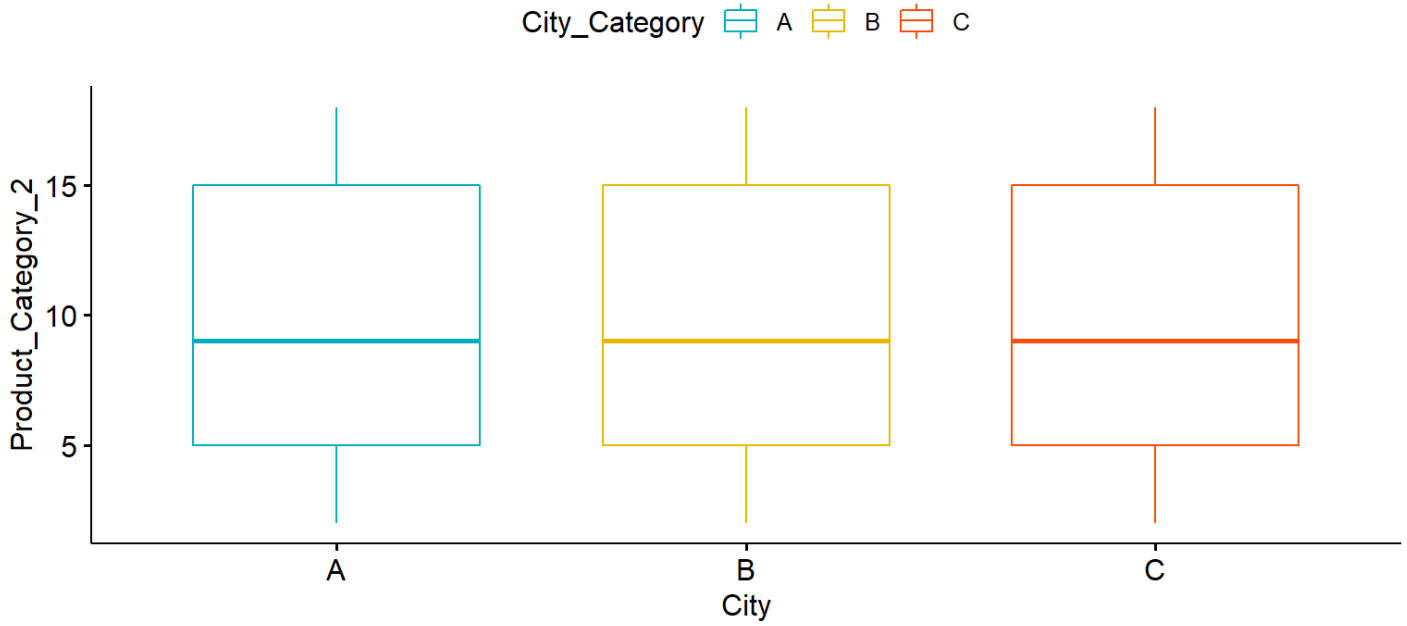
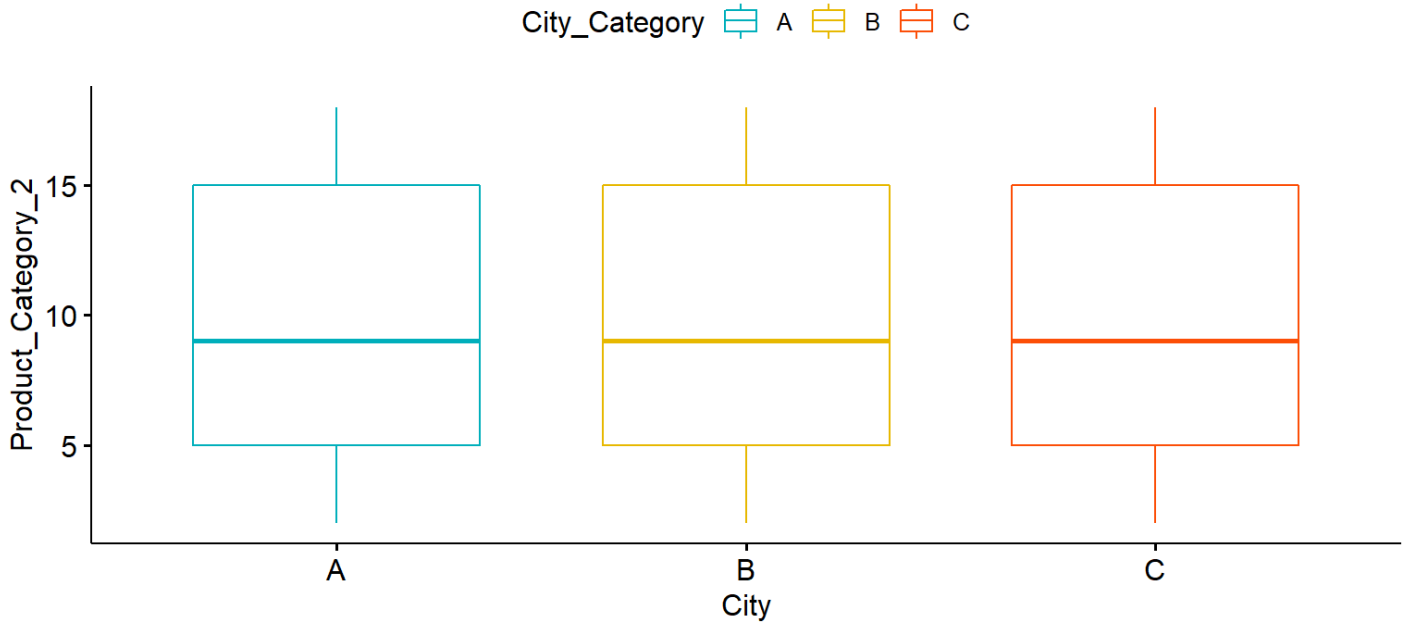


Figure 8 - Boxplot of items in category 3

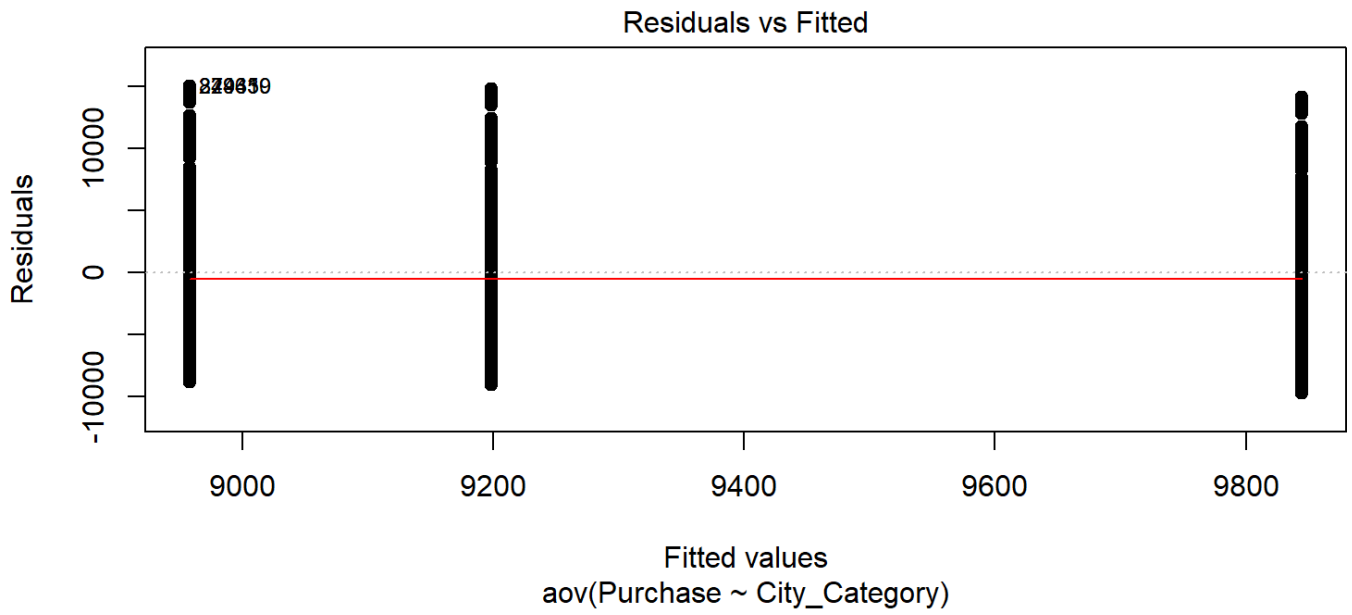


ANOVA analysis

```
## Call:
##    aov(formula = Purchase ~ City_Category, data = friday)
##
## Terms:
##                 City_Category    Residuals
## Sum of Squares   6.796357e+10 1.326961e+13
## Deg. of Freedom             2       537574
##
## Residual standard error: 4968.324
## Estimated effects may be unbalanced
```

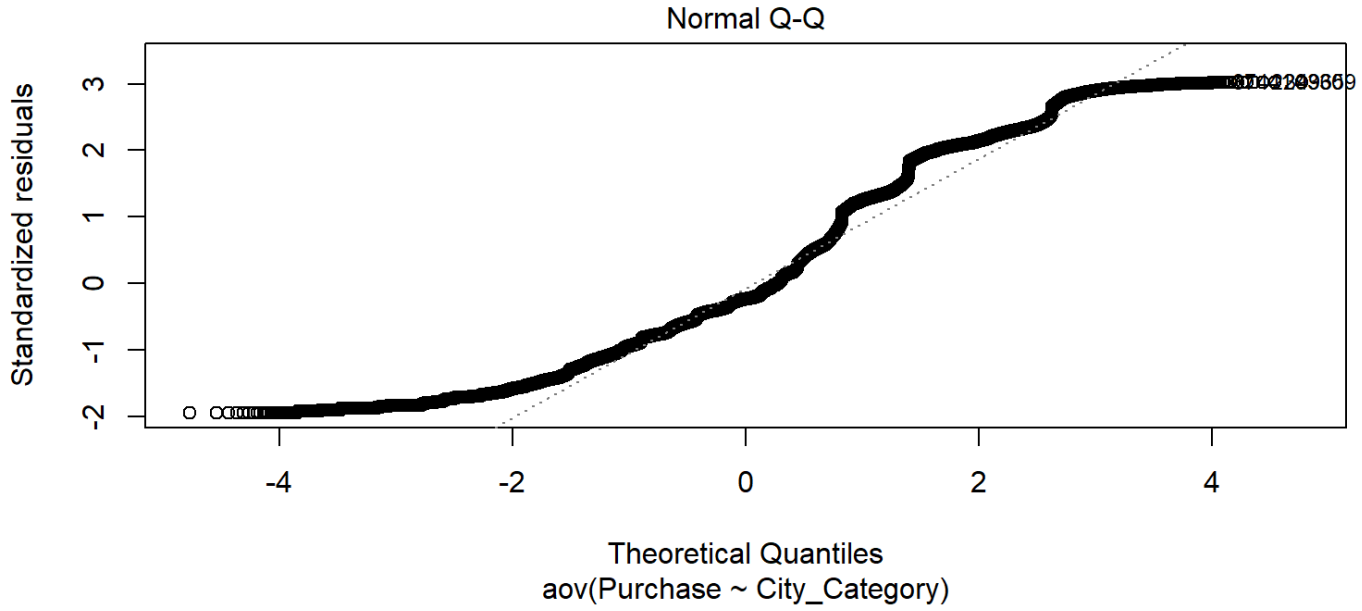
```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Purchase ~ City_Category, data = friday)
##
## $City_Category
##         diff      lwr      upr p adj
## B-A 240.6468 201.4540 279.8397     0
## C-A 886.4308 844.5734 928.2883     0
## C-B 645.7840 608.1907 683.3773     0
```

Figure 9 - check assumption of homogeneity




```
##
##  Bartlett test of homogeneity of variances
##
## data:  Purchase by City_Category
## Bartlett's K-squared = 418.91, df = 2, p-value < 2.2e-16
```

Figure 10 check assumption of normality




```
##
##  Kruskal-Wallis rank sum test
##
## data:  Purchase by City_Category
## Kruskal-Wallis chi-squared = 2766, df = 2, p-value < 2.2e-16
```

```
##                Df    Sum Sq   Mean Sq F value Pr(>F)
## Age             6 6.471e+09 1.078e+09   43.49 <2e-16 ***
## Residuals   537570 1.333e+13 2.480e+07
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Purchase ~ Age, data = friday)
##
## $Age
##                    diff        lwr       upr     p adj
## 18-25-0-17   215.07070   85.20433 344.93707 0.0000216
## 26-35-0-17   294.46209  169.31635 419.60783 0.0000000
## 36-45-0-17   381.35188  252.26796 510.43580 0.0000000
## 46-50-0-17   264.74540  125.10756 404.38324 0.0000005
## 51-55-0-17   600.48974  457.70399 743.27549 0.0000000
## 55+-0-17     433.77170  275.75253 591.79087 0.0000000
## 26-35-18-25   79.39140   22.71707 136.06572 0.0007180
## 36-45-18-25  166.28118  101.37215 231.19022 0.0000000
## 46-50-18-25   49.67470  -34.28511 133.63452 0.5859577
## 51-55-18-25  385.41905  296.32194 474.51615 0.0000000
## 55+-18-25    218.70100  106.80560 330.59641 0.0000002
## 36-45-26-35   86.88979   32.03211 141.74746 0.0000617
## 46-50-26-35  -29.71669 -106.17212  46.73873 0.9137885
## 51-55-26-35  306.02765  223.96380 388.09150 0.0000000
## 55+-26-35    139.30961   32.92933 245.68989 0.0021726
## 46-50-36-45 -116.60648 -199.35088 -33.86208 0.0006458
## 51-55-36-45  219.13786  131.18515 307.09057 0.0000000
## 55+-36-45     52.41982  -58.56652 163.40616 0.8061234
## 51-55-46-50  335.74434  232.92534 438.56334 0.0000000
## 55+-46-50    169.02630   45.92458 292.12803 0.0010140
## 55+-51-55   -166.71804 -293.37932 -40.05676 0.0020083
```

Linear regression

```
##
## Call:
## glm(formula = Purchase ~ City_Category + Age + Gender, data = friday)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
##  -9890   -3448   -1204    2880   15797
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     8050.09      44.15 182.352  < 2e-16 ***
## City_CategoryB   245.86      16.78  14.653  < 2e-16 ***
## City_CategoryC   901.16      18.13  49.702  < 2e-16 ***
## Age18-25         298.60      43.97   6.791 1.11e-11 ***
## Age26-35         408.50      42.45   9.623  < 2e-16 ***
## Age36-45         433.20      43.67   9.919  < 2e-16 ***
## Age46-50         286.98      47.18   6.082 1.19e-09 ***
## Age51-55         602.92      48.26  12.493  < 2e-16 ***
## Age55+           283.58      53.42   5.308 1.11e-07 ***
## GenderM          689.81      15.73  43.867  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 24584010)
##
##     Null deviance: 1.3338e+13  on 537576  degrees of freedom
## Residual deviance: 1.3216e+13  on 537567  degrees of freedom
## AIC: 10673863

##
## Number of Fisher Scoring iterations: 2
```

Logistic regression

```
##      Age  City Gender   Yes    No
## 1   0-17 CityA      M   415   654
## 2  18-25 CityA      M  6957 13954
## 3  26-35 CityA      M 18249 36802
## 4  36-45 CityA      M  6453 12732
## 5  46-50 CityA      M  1763  4488
## 6  51-55 CityA      M  1639  2586
## 7    55+ CityA      M   921  2218
## 8   0-17 CityB      M  1243  2509
## 9  18-25 CityB      M 11373 19625
## 10 26-35 CityB      M 24944 43923
## 11 36-45 CityB      M 12571 23218
## 12 46-50 CityB      M  4971  8656
## 13 51-55 CityB      M  4883  8370
## 14   55+ CityB      M  1437  2276
## 15  0-17 CityC      M  1830  3103
## 16 18-25 CityC      M  8892 12776
## 17 26-35 CityC      M 17437 23987
## 18 36-45 CityC      M 10719 15386
## 19 46-50 CityC      M  4445  7347
## 20 51-55 CityC      M  4116  6390
## 21   55+ CityC      M  3216  5906
## 22  0-17 CityA      F   369  1059
## 23 18-25 CityA      F  1583  4531
## 24 26-35 CityA      F  4628 12369
## 25 36-45 CityA      F  2034  4923
## 26 46-50 CityA      F   357   859
## 27 51-55 CityA      F   453  1291
## 28   55+ CityA      F   118   233
## 29  0-17 CityB      F   483  1053
## 30 18-25 CityB      F  2922  8550
## 31 26-35 CityB      F  5706 15194
## 32 36-45 CityB      F  3260  7556
## 33 46-50 CityB      F  1759  4514
## 34 51-55 CityB      F  1148  3034
## 35   55+ CityB      F   336   979
## 36  0-17 CityC      F   579  1410
## 37 18-25 CityC      F  2110  4361
## 38 26-35 CityC      F  3666  7785
## 39 36-45 CityC      F  2942  5705
## 40 46-50 CityC      F  1712  3655
## 41 51-55 CityC      F  1185  2523
## 42   55+ CityC      F   921  2342
```

```
## , , City = CityA,  = Yes
##
##       Age
## Gender  0-17 18-25 26-35 36-45 46-50 51-55   55+
##      M   415  6957 18249  6453  1763  1639   921
##      F   369  1583  4628  2034   357   453   118
##
## , , City = CityB,  = Yes
##
##       Age
## Gender  0-17 18-25 26-35 36-45 46-50 51-55   55+
##      M  1243 11373 24944 12571  4971  4883  1437
##      F   483  2922  5706  3260  1759  1148   336
##
## , , City = CityC,  = Yes
##
##       Age
## Gender  0-17 18-25 26-35 36-45 46-50 51-55   55+
##      M  1830  8892 17437 10719  4445  4116  3216
##      F   579  2110  3666  2942  1712  1185   921
##
## , , City = CityA,  = No
##
##       Age
## Gender  0-17 18-25 26-35 36-45 46-50 51-55   55+
##      M   654 13954 36802 12732  4488  2586  2218
##      F  1059  4531 12369  4923   859  1291   233
##
## , , City = CityB,  = No
##
##       Age
## Gender  0-17 18-25 26-35 36-45 46-50 51-55   55+
##      M  2509 19625 43923 23218  8656  8370  2276
##      F  1053  8550 15194  7556  4514  3034   979
##
## , , City = CityC,  = No
##
##       Age
## Gender  0-17 18-25 26-35 36-45 46-50 51-55   55+
##      M  3103 12776 23987 15386  7347  6390  5906
##      F  1410  4361  7785  5705  3655  2523  2342
```

```
##
## Call:
## glm(formula = cbind(Yes, No) ~ F1 + F2 + F3, family = binomial,
##     data = dat1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -6.7978  -1.7084   0.0626   2.8174   7.6027
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.109727   0.019087 -58.140  < 2e-16 ***
## F118-25      0.072178   0.018868   3.826 0.000131 ***
## F126-35      0.085793   0.018227   4.707 2.51e-06 ***
## F136-45      0.091187   0.018733   4.868 1.13e-06 ***
## F146-50      0.013568   0.020256   0.670 0.502981
## F151-55      0.090615   0.020642   4.390 1.13e-05 ***
## F155+       -0.066079   0.022951  -2.879 0.003987 **
## F2CityB      0.107888   0.007236  14.909  < 2e-16 ***
## F2CityC      0.304184   0.007719  39.405  < 2e-16 ***
## F3M          0.349060   0.006904  50.560  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4858.4  on 41  degrees of freedom
## Residual deviance:  488.5  on 32  degrees of freedom
## AIC: 893.75
##
## Number of Fisher Scoring iterations: 3
```